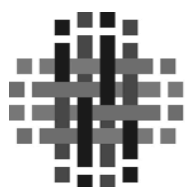


# Corpora, collections, data – reusing outputs of language documentation

Nick Thieberger (University of  
Melbourne)



ARC CENTRE OF EXCELLENCE FOR  
**THE DYNAMICS OF LANGUAGE**

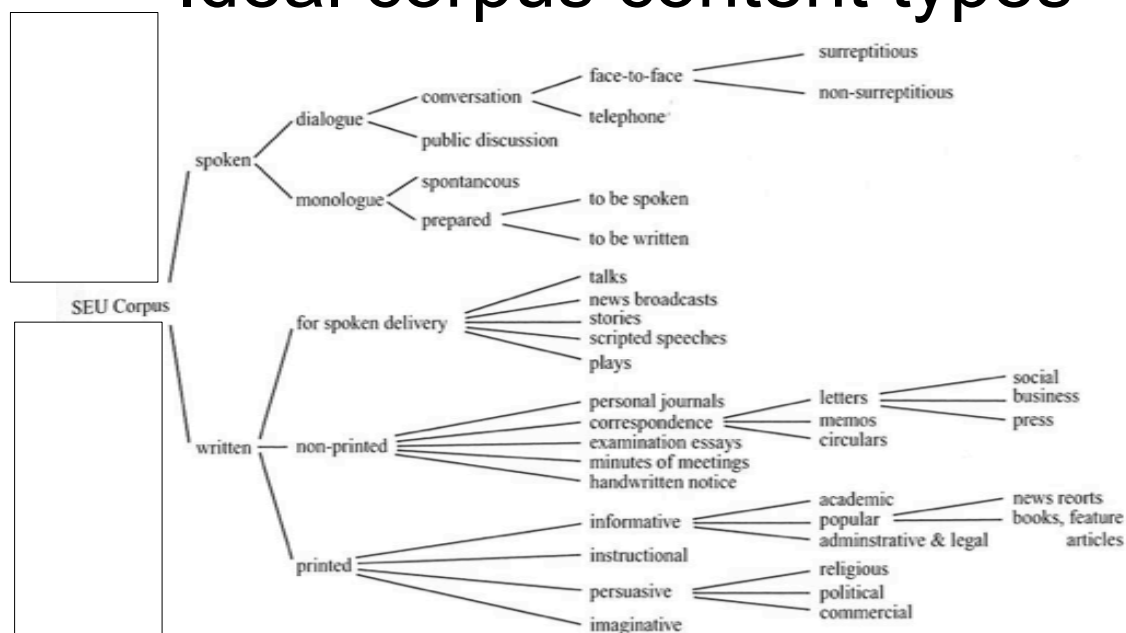
The screenshot shows the website's header with the logo and name, a search bar, and social media links. A navigation bar lists: About | Team | Research | Publications & Seminars | News & Media | Get Involved | What's On | Contact. The main banner features a profile of a man with a network diagram of icons (ears, mouth, brain, etc.) and the text: "Transforming the science of language: how languages vary, how we learn them, how we process them and how they evolve." Below this is a large question in brackets: "Why is language so diverse?". At the bottom of the banner are four tabs: Shape, Learning, Processing, and Evolution. The footer includes a "NEWS AND EVENTS" section with a video player and a link to the archive: <http://www.dynamicsoflanguage.edu.au/> ARCHIVE.

## CoEDL's aims

“Each program will produce varying kinds of material — recorded narratives, conversations, child language records, dictionaries, scholarly articles, statistical analyses and so on — and we aim to archive as much of it as possible. For records of small languages we will create a corpus in which it will be possible to perform the usual corpus operations (search, concordance, collocational search, frequency counts, visualisations based on well-structured data, etc.).

We will also build playable online texts (as in the current EOPAS model) that will ultimately provide stories, with interlinear text and media, for as many languages as we can arrange it for, with a map interface to make it attractive to general users.”

## Ideal corpus content types



- How does a corpus of a small language differ from other corpora?
- What is the ideal form of data for such a corpus?
- What is the most realistic form of data for such a corpus?

## Collections and corpora of small languages

A corpus, for our purposes, is a collection of text(s) that can be searched

But primary materials that are not yet in textual format should also be brought in to the research effort, especially when there is little else available for the language:

-fieldnotes, historical manuscripts, analog recordings

Digitisation: bring analog material into current research methods

# What is a corpus of a small language?

- (A) Constructed in the course of fieldwork, either:
    - incidental to research goal (until recently)
      - typically relies on tools to provide structure
        - if tools were used for transcription, annotation
    - as the main aim of the fieldwork (especially in language documentation)
      - structured ('backslash', .eaf xml, FLEX xml)
  - (B) Built from sources for a particular purpose
    - using tagging and conforming to a particular schema
- not 'balanced' – genres, modes, speakers, ages
  - little choice about what goes in to the corpus
  - everything recorded should be included

# What is a corpus of a small language?

- Ideally
  - audio and video recorded performances
    - narratives
    - procedural texts
    - conversations
    - songs
    - etc

# Actual corpora

- MS Word files on a linguist's computer, maybe intelligible to the creator (often not)
- maybe media recordings, maybe metadata
- maybe transcripts (Elan/CLAN) (XML)
- maybe a structured lexicon (Toolbox/FLEEx) ('backslash, XML)
- maybe structured texts (Toolbox/FLEEx) ('backslash, XML)
- maybe social media in the language (Facebook, Twitter) (unstructured, text)

# Why a corpus?

- What can we know about a language beyond what is given by the grammatical analysis?
- Are there citable example sentences that have context?
- Are there textual examples?
- Is there a set of records referred to by the grammar?

# Corpus linguistics

- well established methods, huge literature
- corpus as a collection of texts, annotated to varying degrees
- however, little representation of 'small' languages in the literature

# Corpus linguistics

A search of the *International Journal of Corpus Linguistics* for 'endangered' or 'indigenous' or 'Aboriginal' or 'First nations' gave zero results.

# Corpus linguistics

O'Keeffe, Anne, and Michael McCarthy, eds. 2010. *The Routledge Handbook of Corpus Linguistics*

'Indigenous' is not an index item, nor is 'endangered' or 'documentation', 'DoBeS', 'Corpo-AfroAs'

'Building a specialised audio-visual corpus' (93-104)  
by Paul Thompson

Does not mention Elan or the work of the DoBeS project, despite appearing in 2010, some 8 years after that project, with a significant emphasis on recording video and annotating multimedia, began.

# Corpus linguistics

Ostler, Nicholas. "Corpora of Less Studied Languages."

In Lüdeling, Anke & Merja Kytö. (eds) 2008.  
*Corpus Linguistics: An International Handbook*.  
Vol. 1. Berlin ; New York: Walter de Gruyter. 457–  
83

## Ostler: *Corpora of Less Studied Languages*

- Lists both corpora and archives
- Discusses an archive as a kind of corpus
- For the Pacific region, he lists the Aboriginal Studies Electronic Data Archive (ASEDA)<sup>†</sup> and the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

- “The modern technology of corpus linguistics allows us to systematically search for particular lexical items and their collocations as well as for constructional patterns and the lexical items they accommodate, to view all findings in a concordance and to analyze the grammatical structures in their natural context.” (154)
- Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In Toshihide Nakayama and Keren Rice (eds) *The Art and Practice of Grammar Writing*, LD&C Special Publication No. 8. pp. 135-157.





# Welcome to CQPweb at the University of Lisbon CORPUS QUERY PROCESSOR

*Please select a corpus from the list below to enter.*

<a href="#">Angolar</a>	<a href="#">Child Corpus</a>	<a href="#">Reference Corpus of Contemporary Portuguese (CRPC) v2.3</a>
<a href="#">Fadambo</a>	<a href="#">Spoken Corpus Mozambique 1986-87</a>	<a href="#">CRPC Portugal only v2.3</a>
<a href="#">Principense</a>	<a href="#">Santome</a>	

Portuguese creole database  
<http://alfclul.clul.ul.pt/CQPweb/>

Menu	Spoken Corpus Mozambique 1986-87: <i>powered by CQPweb</i>	
<b>Corpus queries</b>	<b>Metadata for Spoken Corpus Mozambique 1986-87</b>	
Standard query	Corpus name	Spoken Corpus Mozambique 1986-87
Restricted query	CQPweb's short handles for this corpus	moz / MOZ
Word lookup	Total number of corpus texts	48
Frequency lists	Total words in all corpus texts	140,198
<b>Corpus info</b>	Word types in the corpus	7,888
View corpus metadata	Type:token ratio	0.06 types per token
Corpus documentation	<b>Text metadata and word-level annotation</b>	
<b>About CQPweb</b>	The database stores the following information for each text in the corpus:	There is no text-level metadata for this corpus.
CQPweb main menu	The <b>primary</b> classification of texts is based on:	A primary classification scheme for texts has not been set.
CQPweb manual	Words in this corpus are annotated with:	lemma
Who did it?		pos
Latest news	The <b>primary</b> tagging scheme is:	pos
	Further information about this corpus is available on the web at:	<a href="http://alfclul.clul.ul.pt/CQPweb/doc/mozambique.html">http://alfclul.clul.ul.pt/CQPweb/doc/mozambique.html</a>

<http://alfclul.clul.ul.pt/CQPweb/moz/>

# Let 100 flowers bloom

Recognise that corpora will be built in various ways and then try to convert results into a standard format

ANNIS - Integrating Annotations from Different Tools and Tag Sets

Chiarcos, Christian, Stefanie Dipper, Michael Balick, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. n.d. "A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets." *Traitement Automatique Des Langues* 49 (2/2008): 217–46.

## Bringing corpora into a standard operating environment

- Legacy material needs conversion
- We can provide templates for new research
- Train new researchers to use the templates
- Build 'workbenches' that sit over standard data sets to create useful search and visualisation

## Using finite resources to either make many records locatable or to enrich a small set of records for automated access

- How much detail to use to annotate a set of material?
- Example of two approaches:
  - just metadata to make page images locatable
  - textual content to make it all searchable

## Presenting fieldnotes as images

- Case study: Arthur Capell's papers >14,000 pages
  - Paper originals stored in the National Library of Australia
  - Simple means of providing access to unique resources
  - No time or funding to encode the text of the images
- Metadata created by the executor of the estate
- Photographic images (not scanned)
  - Portable photographic rig, images taken in the executor's house
  - Rapid and easy capture

PARADISEC

Home  
Provenance  
Series List  
Search

Pacific and Regional Archive for Digital Sources in Endangered Cultures

## Arthur Capell (1902-1986), Guide to Records

Listed by Peter Newton, Austehc and PARADISEC

Arthur Capell was an Australian linguist and ethnographer who spent much time recording and documenting both Australian Aboriginal languages and endangered languages in the Asia - Pacific region. The Arthur Capell textual collection consists of some thirty boxes of notes, transcripts, and other supporting materials. What is documented in this guide are some of the records relating to his non-Australian linguistic work that have been digitally imaged.

**Please do not copy material** from this site for further distribution but rather link to this site. PARADISEC has raised funds to digitise this collection and would like to be recognised for the work that we have put into developing the online presentation of fieldnotes. If you copy and distribute this data and do not acknowledge PARADISEC's work then we will have to put password protection on the data.

**Copyright:** ParadiSec believes that many of the items provided through this guide are no longer the subject of copyright restrictions, or have been cleared for display in this service by the Copyright owners. However, ParadiSec invites any individuals who believe they hold current rights over items provided through this service to make contact.

[\[Details\]](#)

- [About the records](#)
  - [Scope and content](#)
  - [How to use this finding aid](#)
  - [Archival terms](#)
- [Provenance - creators and former custodians](#)
- [Series list and summary descriptions](#)
- [Inventory listing by series](#)
  - Series 01 - [Personal and Biographical Material](#)
  - Series 02 - [General Linguistic and Ethnological Materials](#)
  - Series 03 - [Indonesia - Regions and Languages](#)
  - Series 04 - [Bougainville, Buka and Offshore Islands - Regions and Languages](#)
  - Series 05 - [Melanesia and Oceania - Regions and Languages](#)
  - Series 06 - [Solomon Islands - Regions and Languages](#)
  - Series 07 - [Timor - Regions and Languages](#)
  - Series 08 - [Vanuatu - Regions and Languages](#)
  - Series 09 - [Fiji - Regions and Languages](#)

**Arthur Capell (1902-1986), Guide to Records**

Item: VEFAT25

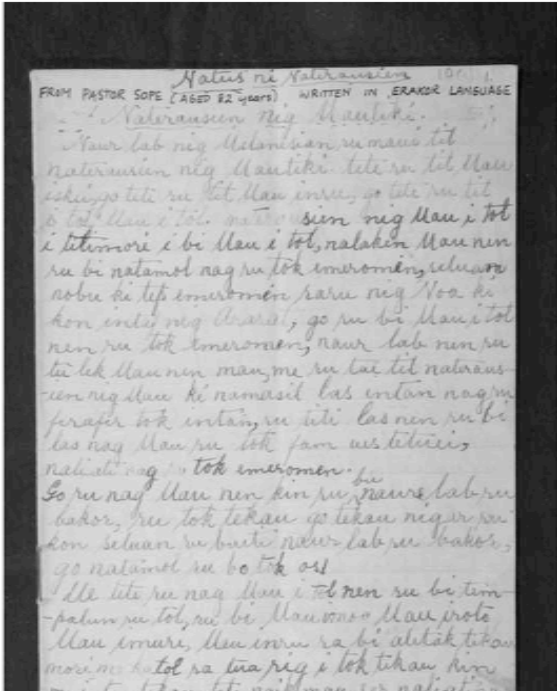
Contact PARADISEC

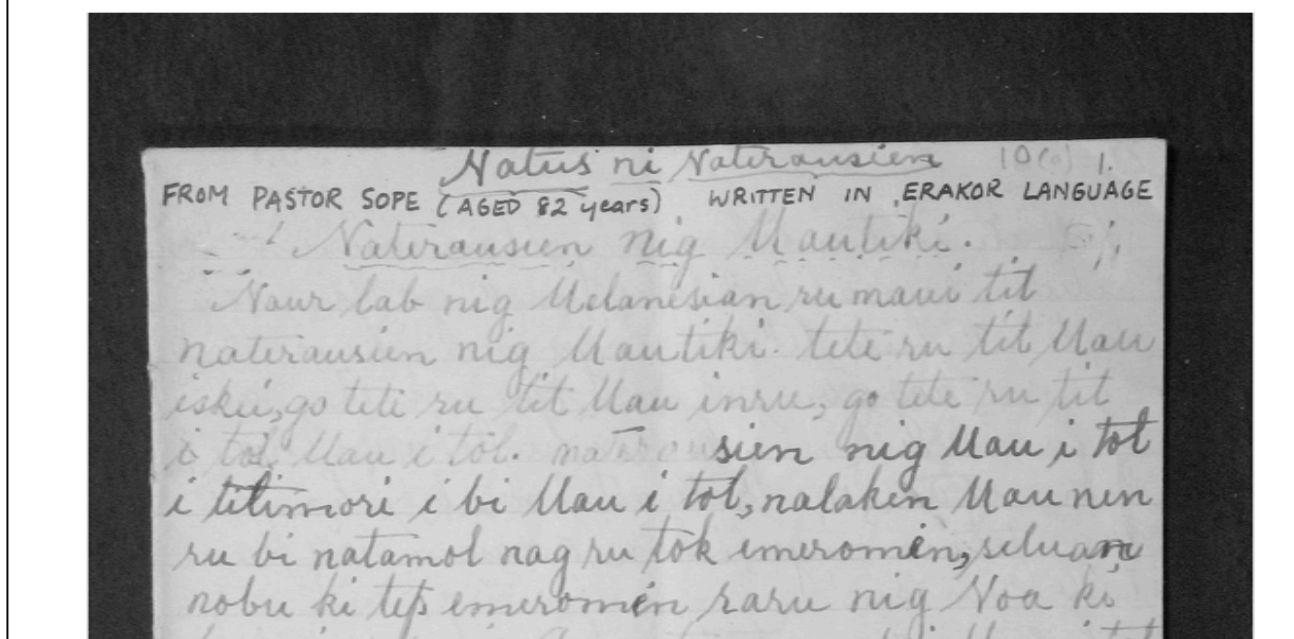
To print this page you may need to select landscape paper orientation for better results.

Read
Enlarge

<-- Image 1 of 30 -->

Back to Guide





<http://paradisec.org.au/fieldnotes/AC2.htm>

*Natus ni Natrausien* 1000 1.  
FROM PASTOR SOPE (AGED 82 years) WRITTEN IN ERAKOR LANGUAGE  
*Natrausien nig Mautiki.*  
Naur lab nig Melanesian ru mau i tit  
natrausien nig Mautiki. tete ru tit Mau  
iskei, go tete ru tit Mau inru, go tete ru tit  
i tol, Mau i tol. natrausien nig Mau i tol  
i titimori i bi Mau i tol, nalaken Mau nen  
ru bi natamol nag ru tok emeromen, seluan  
nobe ki tep emeromen raru nig. Noa ki

#### (1) Natrausien nig Mautiki

1. Naur lab nig Melanesian ru mau til natrausien nig Mautiki tete ru til Mau iskei, go tete ru til Mau inru, go tete ru til Mau i tol i titimori i bi Mau i tol, nalaken Mau nen ru bi natamol nag ru tok emeromen, seluan nobe ki tep emeromen raru nig Noa ki kon intaf nig Ararat, go ru bi Mau i tol nen ru tok emeromen, naur lab nen ru tu lek Mau nen mau, me ru tae til natrausien nig Mau ki namasil las intan nagru ferafer tok intan, ru tili las nen ri bi las nag Mau ru tok fam uis tetuei, naliati nag ru tok emeromen.

2. Go ru nag Mau nen kin ru bu naure lab ru bakor, ru tok tekau go tekau nigar ru kon seluan ru bueti naur lab ru bakor, go natamol ru bo tok os!

3. Me tete ru nag Mau i tol nen ru bi tempalun ru tol, ru bi Mau imoo Mau iroto Mau imuri, Mau inru ra bi atlak tikau mori me katol ra tua prig i tok tikau kin me i tu tkau tete naik mau ser naliatinag ru ban tekau nigar

#### (1) Natrausien nig Mautiki

1. Naur lap ni Melanesia rumau til natrausien ni Mau. Tete ru til Mau iskei, tete rutil inru, go tete rutil Mau itol. Natrausien ni Mau itol itilmori kin ipi Mau itol, nalaken Mau nen ru pu natariol nag ru tok emeromen. Selwan nab itap emeromen go raru ni Noa ikon ntaf ni Ararat, naur lap rutap lek Mau nen mau me rutae til natrausien ni Mau ki nmasil las nag rufrafer tok ntan, rutli nag las nen rupi las nag Mau ruto fam wes tetuei, malnen ruto emeromen.

2. Go runag Mau nen kin ru po naur lap rupakor. Ruto pan tkau go tkau negar rukon go selwan rupueti naur lap rupakor, go natariol ru po tkos.

3. Me tete runag Mau tol nen rupi temp alun, rupi Mau Imoo, Mau Iroto, ga Mau Imuri. Mau inru rapi atlak tkau mori me katol ratua tkau gar ito tkau kin me naliati itap tkau tete naik mau ser naliati nag rupan tkau gar rato tua, tete naik nag gar ratkaus.

#### (1) Storian blong Mautiki

1. Fulap aelan blong Melanesian oli talem stori blong strong man blong faet. Sam oli talem se igat wan, sam oli talem igat tu ma sam oli talem se igat tri. Stori blong trifala strong man blong faet hemi tru from ol man blong faet ya oli ol man we oli stap long wol. Taem we wota i bin drae long wol mo Bot blong Noa i fas long hil blong Ararat, fulap aelans oli no luk ol man blong faet ya be oli save talem stori blong hem from ol pispis klas we oli stap olbaot oli se hemi ol kap we ol man ya oli stap kakae long hem bifo taem oli stap long wol.

2. Mo oli talem se ol strong man blong faet ya nao oli pulum ol plante aelan oli kamaot. Oli stap go huk mo huk blong olgeta i fas mo taem oli pulum, plante aelan oli kamaot, mo ol man oli stap long hem.

3. Be sam oli talem se ol trifala strong man blong faet ya oli brata, ol nem blong olgeta oli Imoo, Iroto mo Imuri. Tu long trifala strong man ya tufala igat huk be tufala i givim huk blong tufala



## Item details

<input type="checkbox"/> Private: Hide metadata from all users	
<b>Item ID</b>	NT3-storian (Collection Details)
<b>Title</b>	Storian Blong Pastor Sope long lanwis blong Saot Efate we oli bin kamaot samples long yia 1950
<b>Description</b>	<p>Preface -- The stories here were written by Pastor Sope in the 1950s. There were another two stories in English which I haven't included here. Pastor Sope wrote these stories in the old language of South Efate . I found these stories in Arthur Capell's materials which were held by Peter Newton at his house in Balmain. Thanks to Peter Newton for looking after these papers (see <a href="http://catalog.paradisec.org.au/collections/AC2/items/VEFAT25">http://catalog.paradisec.org.au/collections/AC2/items/VEFAT25</a>). Dina Thieberger typed the stories and Endis Kalsarap translated them into modern South Efate and into Bislama. -- Fes toktok -- Ol storian ia oli kamaot long sam pepa we Pastor Sope i bin raetem samples long yia 1950. -- Pastor Sope ibin raetem storian ia, mo igat tufala storian mo long Inglis we mifala i no bin joenem long plesia. -- Ol storian we i stap long smol buk ia Pastor Sope i bin raetem long olfala lanwis blong bifo. -- Mi bin faenem ol storian ia taem mi bin</p>

## Content Files (20)

Filename ▲▼	Type ▲▼
NT3-storian-019.jpg	image/jpeg
NT3-storian-020.jpg	image/jpeg
NT3-storian-021.jpg	image/jpeg
NT3-storian-022.jpg	image/jpeg
NT3-storian-023.jpg	image/jpeg
NT3-storian-024.jpg	image/jpeg
NT3-storian-025.jpg	image/jpeg
NT3-storian-026.jpg	image/jpeg
NT3-storian-027.jpg	image/jpeg
NT3-storian-028.jpg	image/jpeg
NT3-storian-029.jpg	image/jpeg
NT3-storian-030.jpg	image/jpeg
NT3-storian-031.jpg	image/jpeg
NT3-storian-032.jpg	image/jpeg
NT3-storian-033.jpg	image/jpeg
NT3-storian-034.jpg	image/jpeg
NT3-storian-035.jpg	image/jpeg
NT3-storian-036.jpg	image/jpeg
NT3-storian-storian.pdf	application/pdf
NT3-storian-storian.txt	text/plain

## One-off examples of encoded text for a small language

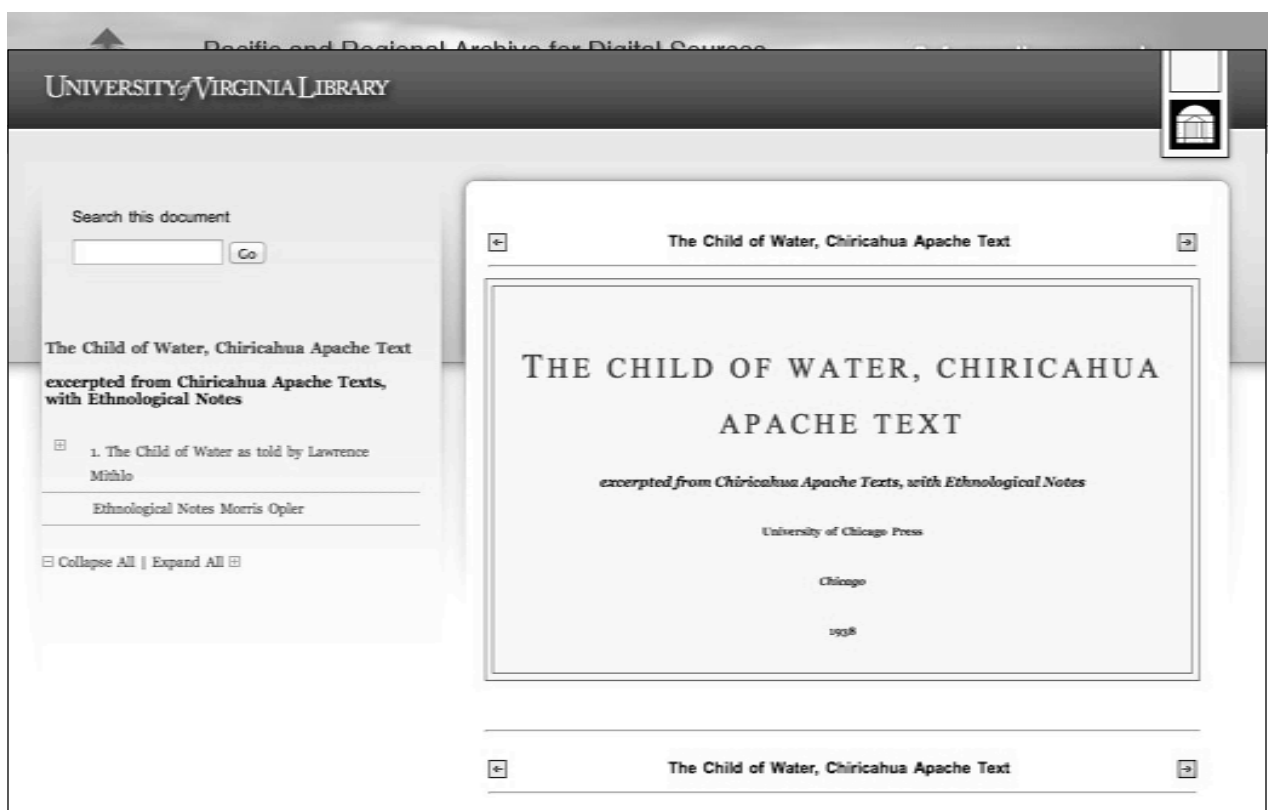
Web-accessible Apache language linguistic database and text archive, a republication of Hoijer's 1938 monograph originally published by the University of Chicago.

<http://etext.lib.virginia.edu/apache/ChiMesc2.html> (2008)

now resolves to ->

<http://xtf.lib.virginia.edu/xtf/view?docId=Apache/uvaGenText/tei/Chi01.xml>

1. 55 Apache language texts (in Chiricahua and Mescalero dialects) representing a variety of genres: stories, songs, prayers and speeches, performed by nine different Apache speakers
2. English translations of the Apache texts, developed by Hoijer in consultation with Apache speakers
3. Ethnological notes to the texts, contributed by Morris Opler, a cultural anthropologist
4. Hoijer's linguistic notes, containing morphological analysis of words and phrases as they appear in the texts
5. Hoijer's grammatical outline of Chiricahua and Mescalero Apache language varieties
6. a search-engine for Apache and English language searches
7. a lexicon to the texts developed from Hoijer's linguistic analysis



Pacific and Regional Archive for Digital Sources

Search this document

The Child of Water, Chiricahua Apache Text  
excerpted from Chiricahua Apache Texts,  
with Ethnological Notes

1. The Child of Water as told by Lawrence Mithlo

1.1. At the beginning the Creator

1.2. Afterwards Child of the Water was born.

1.3. So there were four existing at the very beginning [of time].

1.4. And Child of the Water was the child of White Painted Woman.

1.5. He who is called Giant

1.6. Then White Painted Woman went about weeping.

1.7. Then White Painted Woman prayed.

1.8. Then:

1.9. "What shall I do that this baby of mine is [safely] reared?"

1.10. Every day he who is called Giant customarily came to her.

The Child of Water, Chiricahua Apache Text

1. The Child of Water as told by Lawrence Mithlo

At the beginning the Creator

(1.1)<sup>[1]</sup>  
*ʔitsésʔj Bík'ehgo'wíndát góǵínd'a.*  
*Dájik'eh bédooǵísi.*  
*'Ákoo ʔsdzónádeeshé 'íidj góǵínd'a.*

At the beginning the Creator<sup>[1]</sup> existed.  
 Everyone knows about him.  
 And White Painted Woman<sup>[2]</sup> also existed.

(1.1) Linguistic Notes

1. 'ʔitsésʔj' 'at the beginning'. *ʔitsé* 'the first, the beginning, first' [part.], *-shj* 'from' [pp.]

2. *Bík'ehgo'wíndát* 'the Creator'. A compound of *bík'ehgo* 'being by reason of him', *wíndát* 'there is life', and *-át* relative enclitic referring to persons [see Grammatical Sketch, §20]. *bík'eh* 'his spiritual power, in his charge, by reason of him', a noun with third person possessive pronoun; *-go* subordinating enclitic. *ʔin dót* 'there is life, life's verb [apparently imp. neut. and with prefix ʔ-] found only in the third person.

3. *góǵínd'a* 'he lived, it is said'. *góǵí*, third person of: *go-ní-...-l'* 'to live, to exist' [imp. neut. intr.]. The prefix *go-* is often found as a derivational prefix but it is difficult to isolate its meaning. Cf. *níj* 'he is' and *ndéníj* 'he is a man' [ *ndé* 'man' ]. *ní-* is a prefix found with verbs defining adjectival notions. In the third person, it disappears leaving a high tone on the vowel of the preceding



Pacific and Regional Archive for Digital Sources  
in Endangered Cultures

Safeguarding research  
in Australia's region

## Wide shallow or narrow deep

- Where effort is being made to annotate and enrich a corpus that should then be usable within a cross-corpus search
- Where the effort is to make primary material available, that too should be available, even at the level of metadata
  - with an annotation framework that permits it to be enriched



# EOPAS – text and media corpus presentation

Upload text in interlinear format

Upload media

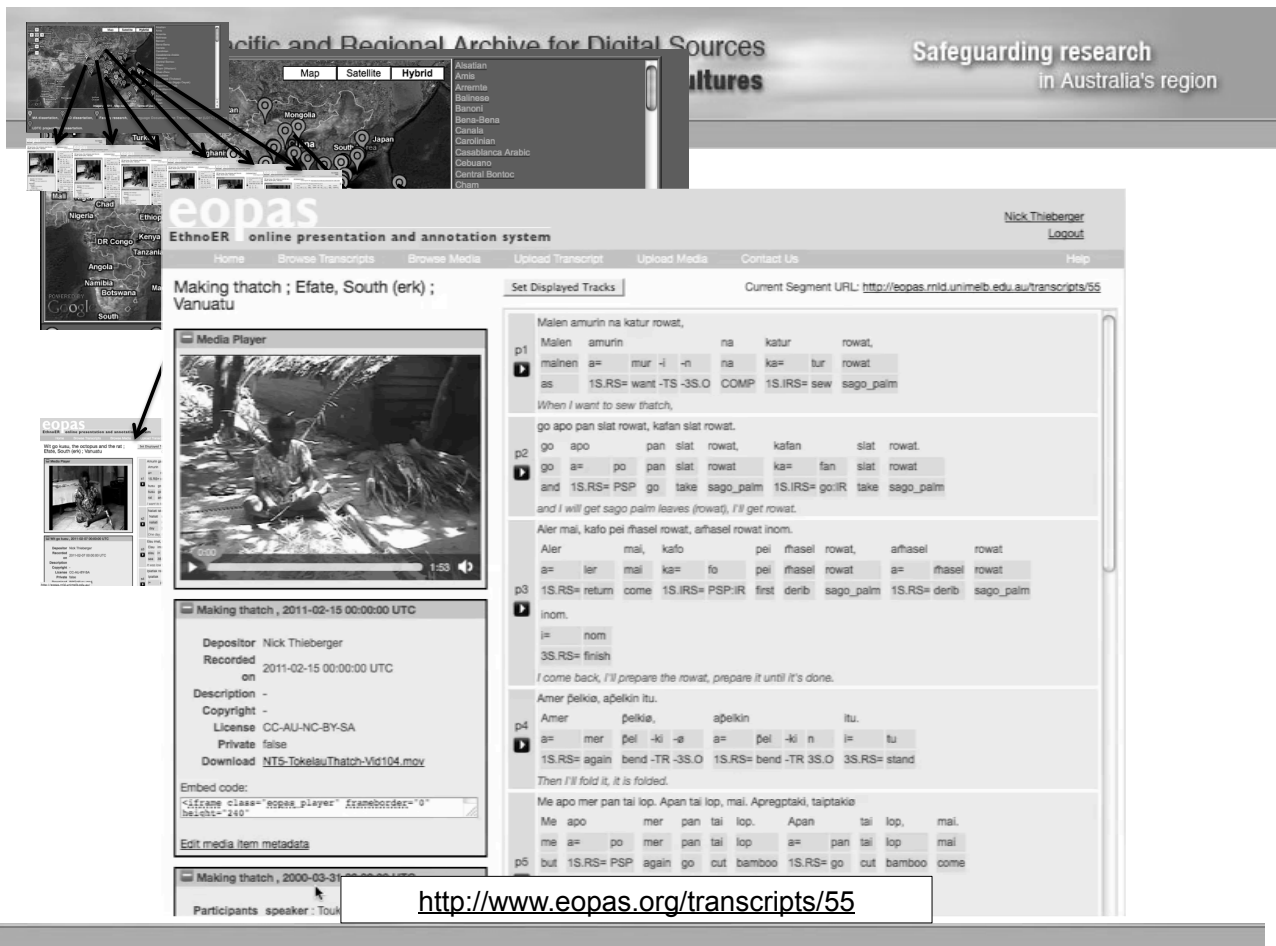
Present text and media online

Citable text and media (to the level of the morpheme)

- Planned

Allow media in PARADISEC to be called directly

Provide online annotation



**EOPAS** online presentation and annotation system

Home Browse Transcripts Browse Media Upload Transcript Upload Media Contact Us Help

Current Segment URL: <http://eopas.mil.unimelb.edu.au/transcripts/55>

**Making thatch ; Efate, South (erk) ; Vanuatu**

Media Player

Making thatch , 2011-02-15 00:00:00 UTC

Depositor: Nick Thieberger  
Recorded on: 2011-02-15 00:00:00 UTC  
Description: -  
Copyright: -  
License: CC-AU-NC-BY-SA  
Private: false  
Download: [NTS-TokelauThatch-Vid104.mov](#)

Embed code:  
`<iframe class="eopas_player" frameborder="0" height="240">`

Edit media item metadata

Participants speaker: Tok

Set Displayed Tracks

Malen amurin na katur rowat,  
p1 Malen amurin na katur rowat,  
mainen a= mur -i -n na ka= tur rowat  
as 1S.RS= want -TS -3S.O COMP 1S.IRS= sew sago\_palm  
When I want to sew thatch,  
go apo pan slat rowat, kafan slat rowat.  
p2 go apo pan slat rowat, kafan slat rowat.  
go a= po pan slat rowat ka= fan slat rowat  
and 1S.RS= PSP go take sago\_palm 1S.IRS= go:IR take sago\_palm  
and I will get sago palm leaves (rowat), I'll get rowat.  
Aler mai, kafo pei rhasel rowat, aphasel rowat inom.  
p3 Aler mai, kafo pei rhasel rowat, aphasel rowat  
a= ler mai ka= fo pei rhasel rowat a= rhasel rowat  
1S.RS= return come 1S.IRS= PSP:IR first derib sago\_palm 1S.RS= derib sago\_palm  
inom.  
i= nom  
3S.RS= finish  
I come back, I'll prepare the rowat, prepare it until it's done.  
Amer pelkie, apelkin itu.  
p4 Amer pelkie, apelkin itu.  
a= mer pel -ki -e a= pel -ki n i= tu  
1S.RS= again bend -TR -3S.O 1S.RS= bend -TR 3S.O 3S.RS= stand  
Then I'll fold it, it is folded.  
Me apo mer pan tai lop. Apan tai lop, mai. Apreptaki, taipakie  
Me apo mer pan tai lop. Apan tai lop, mai.  
p5 me a= po mer pan tai lop a= pan tai lop mai  
but 1S.RS= PSP again go cut bamboo 1S.RS= go cut bamboo come

<http://www.eopas.org/transcripts/55>



<http://www.eopas.org/transcripts/55>

**Concordance**

Track	Track Name	Track Content
mainen a=	mur	-i-n na ka= tur rowat
a=	mur	-i-n gaag puserak N...
go a=	mur	-i-n na ka= trau...
teetwai	mur	pa= sol nalanen knen
...mat top	mur	nag= siwer ur nskau
i= tu ft lefak	mur	-i-n na
me ag ku= ku ku=	mur	-i-n na ku= fak nan...
...ai go maarik wan ki=	mur	-i-n nen kin ke= fa...
...n ki -e	mur	-i-n na ka= welu ag
tu= tae pan a=	mur	-i-n na tu= fak elau
...u	mur	-i-n na
kusu	mur	-i-n na
...e	mur	-i-n na
go kusu	mur	-i-n na
np?au wit go ki= mer	mur	-i-n na a me ku=
np?au wit go ki= mer	mur	-i-n na a me ku=
na E	mur	-i-n na
kinou a= to	mur	-i-n na
i= to	mur	-i-n na
...o esan to pan pan	mur	-i-n na
i= na a= to	mur	-i-n na
...tu	mur	-i-n na

**Set Displayed Tracks**

Current Segment URL: <http://www.eopas.org/transcripts/69#/p3/w6/m2>

**Transcript 69#**

Mainen ina kefak, itu sa imur na kefak Ermag.

p3 Mainen ina kefak, itu sa imur na kefak Ermag.

mainen i= na ke= fak i= tu sa i= mur na ke= fak Ermag

as 3S.RS= want 3S.IRS= go\_to IR 3S.RS= give here 3S.RS= want say 3S.IRS= go\_to IR p.name

When he wanted to go, he was there and he wanted to go to Eromango.

Mainen isiwur ur ntas kin ipak Ermag, go ntas ipathor na?utwen.

p4 Mainen isiwur ur ntas kin ipak Ermag, go ntas ipathor na?utwen.

mainen i= siwer ur ntas kin i= pak Ermag go ntas i= parthor na?utwen

as 3S.RS= walk follow saltwater COMP 3S.RS= go\_to Eromango and saltwater 3S.RS= find knee -V 3S.DP

When he crossed the sea to Eromango, the sea came to his knee.

Esan mana ruta lom mau.

p5 Esan mana ruta lom mau.

esan mana ru= ta lom mau

place group 3P.RS= not wet NEG2

Here (indicating his chest) wasn't wet.

Me imai pak Ermag pan kaimer ler mai go naliati iskei welkia Ermag, lpi, kutae to Efet go kuto lek Ermag.

p6 Me imai pak Ermag pan kaimer ler mai go naliati iskei welkia Ermag, lpi, kutae to Efet go

me i= mai pak Ermag pan kaimer ler mai go naliati i= skel welkia Ermag i= pi ku= tae to Efet go

and 3S.RS= come to p.name go ES= again return come and day 3S.RS= one thus p.name 3S.RS= be you-know stay Efet and

p7 kuto lek Ermag

ku= to lek Ermag

2S.RS= PROG look p.name

He went to Eromango and he came back, and one day, well, Eromango it was, you could be on Efet and you could see Eromango.

Eheltig rhas.

p7 Eheltig rhas.

e= eheltig rhas

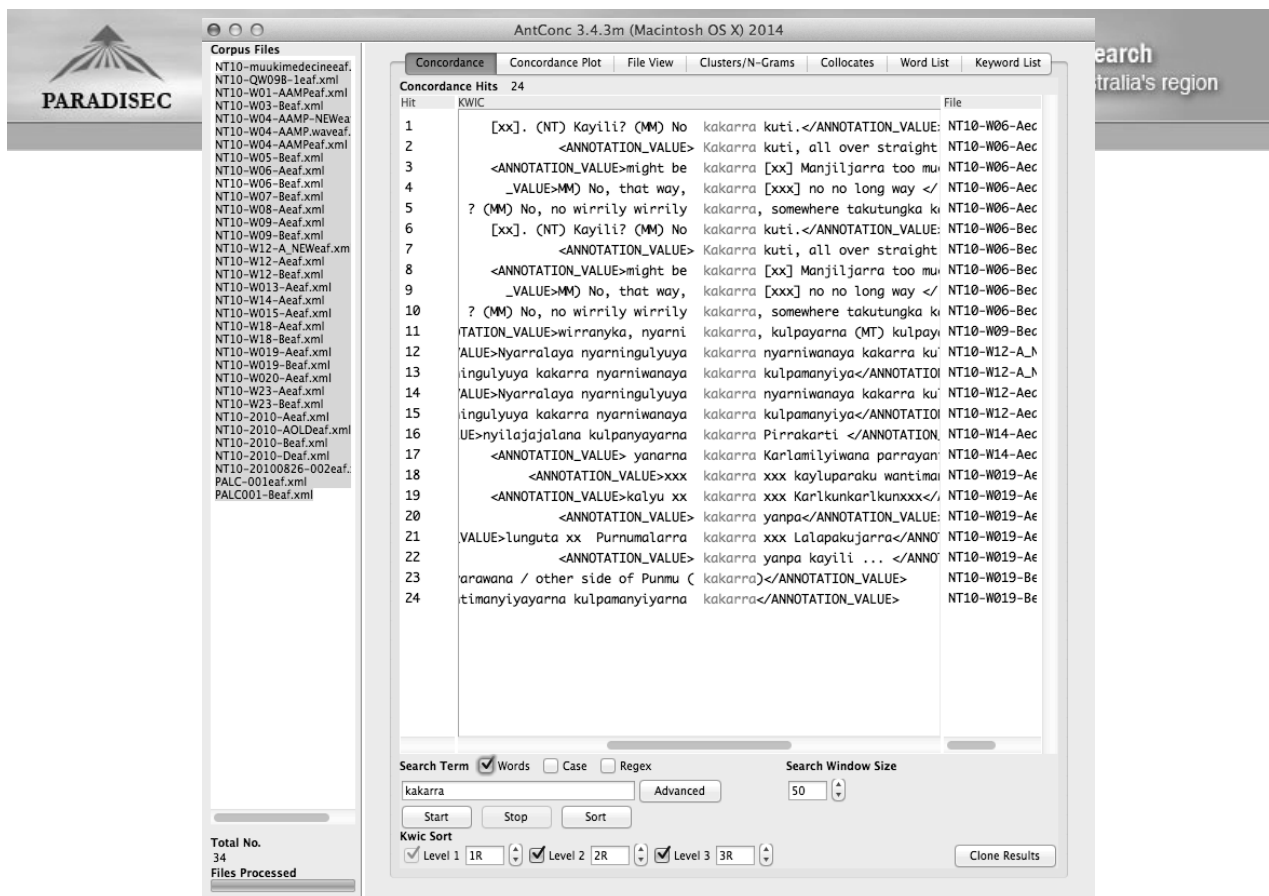
LOC= close only

<http://www.eopas.org/transcripts/69#l/p3/w6/m2>


# Using AntConc to query small corpora

- AntConc – won't open .eaf files
- Knows about xml, so rename .eaf as .xml

<http://www.laurenceanthony.net/software/antconc/>



The screenshot shows the AntConc 3.4.3m (Macintosh OS X) 2014 interface. The 'Corpus Files' list on the left includes files like NT10-muukimedecineeaf.xml, NT10-QW09B-1eaf.xml, and others. The main window displays a concordance search for the term 'kakarra'. The search results are shown in a table with columns for Hit, KWIC, and File. The search term is 'kakarra', and the search window size is set to 50. The search results show various occurrences of 'kakarra' in the corpus, including phrases like 'Kakarra kuti, all over straight' and 'Kakarra [xxx] Manjiljarra too mu'.



AntConc 3.4.3m (Macintosh OS X) 2014

Concordance

Concordance Plots

File View

Clusters/N-Gram

Collocate

Word List

Keyword List

Total No. of Cluster Types

25

Total No. of Cluster Tokens

35

Rank	Freq	Range	Cluster
1	4	2	kalyu wirranykarti
2	2	2	kalyu nganaku
3	2	2	kalyu wiranykarti
4	2	2	kalyu wirranpaya
5	2	2	kalyu wirranyia
6	2	2	kalyu wirranyja
7	2	2	kalyu wirranyjaya
8	2	2	kalyu</annotation
9	1	1	kalyu .. kayili
10	1	1	kalyu [xxwindnoise
11	1	1	kalyu [xxx
12	1	1	kalyu kalyuyarna
13	1	1	kalyu kuwiyi
14	1	1	kalyu pa
15	1	1	kalyu parnijanu
16	1	1	kalyu wangalinya
17	1	1	kalyu wirranykarti
18	1	1	kalyu xx
19	1	1	kalyu yala
20	1	1	kalyu yalangulyu
21	1	1	kalyu yantulja
22	1	1	kalyu, come
23	1	1	kalyu, jutupa
24	1	1	kalyu, manjiljarra
25	1	1	kalyu]</annotation

Search Term

☒ Words

☐ Case

☐ Regex

☐ N-Grams

Cluster Size

Min. 2

Max. 2

Start

Stop

Sort

Sort by

☐ Invert Order

Search Term Position

☒ On Left

☐ On Right

Clone Results


Total No.

34

Files Processed

ling research

in Australia's region



AntConc 3.4.3m (Macintosh OS X) 2014

Concordance

Concordance Plots

File View

Clusters/N-Gram

Collocates

Word List

Keyword List

File View Hits

171

File

Wubuy-Lexicon.xml

</Astr>

</LexEntry\_HeadWord>

<LexEntry\_Senses>

<LexSense number="1" id="hvo3753">

<MoMorphSynAnalysisLink\_MLPartOfSpeech>

<Astr ws="en">

<Run ws="en">n</Run>

</Astr>

</MoMorphSynAnalysisLink\_MLPartOfSpeech>

<MoMorphSynAnalysisLink\_MLInflexionClass>

<Astr ws="en">

<Run ws="en">fem</Run>

</Astr>

</MoMorphSynAnalysisLink\_MLInflexionClass>

<LexSense\_Definition>

<Astr ws="en">

<Run ws="en">hawksbill turtle, Eretmochelys imbricata

(has large, parrotlike beak Syn: garruba. , a fly sp. (uncommon

sense).</Run>

</Astr>

</LexSense\_Definition>

</LexSense>

</LexEntry\_Senses>

Search Term

☒ Words

☐ Case

☐ Regex

Hit Location

fem

Advanced

9

Start

Stop

Clone Results

Total No.

36

Files Processed

ling research

in Australia's region

## Conclusions

- Corpora of small languages typically vary considerably
- Best to address each in its own terms
- Build methods to operate with this diversity
- Provide templates for new researchers

## References

- Chiarcos, Christian, Stefanie Dipper, Michael Balick, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. n.d. "A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets." *Traitement Automatique Des Langues* 49 (2/2008): 217–46.
- Himmelfmann, Nikolaus P. 2012. Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation & Conservation* 6. 187-207.
- Lüdeling, Anke, and Merja Kytö, eds. 2008. *Corpus Linguistics: An International Handbook*. Vol. 1. 2 vols. W. de Gruyter.
- Lüdeling, Anke, and Merja Kytö, eds. 2009. *Corpus Linguistics: An International Handbook*. Vol. 2. 2 vols. Handbücher Zur Sprach- Und Kommunikationswissenschaft = Handbooks of Linguistics and Communication Science, Bd. 29.1-. Berlin; New York: Walter de Gruyter.
- Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In Toshihide Nakayama and Keren Rice (eds) *The Art and Practice of Grammar Writing*, LD&C Special Publication No. 8. pp. 135-157.
- Paul Thompson. 2010. Building a specialised audio-visual corpus. In . O'Keeffe, Anne, and Michael McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. 1st ed. 93-104. Routledge Handbooks in Applied Linguistics. New York, N.Y: Routledge.
- Woodbury, Anthony. 2014. Archives and audiences: toward making endangered language documentations people can read, use, understand, and admire. In David Nathan and Peter Austin (eds.) *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*. London: SOAS. 19-36.
- Wynne, Martin, ed. 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oakville, CT: Oxbow Books.

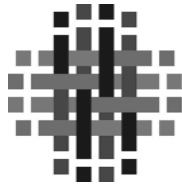


Pacific and Regional Archive for Digital Sources  
**in Endangered Cultures**

**Safeguarding research**  
in Australia's region



Australian Research Council – LIEF grants LE110100142,  
LE0560711, LE0453247  
ARC DP0450342, DP0984419, FT140100214



ARC CENTRE OF EXCELLENCE FOR  
**THE DYNAMICS OF LANGUAGE**